UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI

"M. FANNO"

CORSO DI LAUREA IN ECONOMIA E MANAGEMENT

PROVA FINALE

"STIMA DEI RENDIMENTI DELL'ISTRUZIONE: UTILIZZO DEL CREDO RELIGIOSO COME VARIABILE STRUMENTALE E IL PROBLEMA DELLA SUA RILEVANZA"

RELATORE:

CH.MO PROF. NUNZIO CAPPUCCIO

LAUREANDO: DANIELE COLOMBO

MATRICOLA N. 1066280

ANNO ACCADEMICO 2015-2016

Parole: 11750

Sommario

Introduzione

La differenza nei redditi percepiti da lavoratori diversi può essere spiegata in molti modi. L'obiettivo di questa tesi è, nella prima parte, di proporre un modello economico che tenti di spiegare questa situazione, nelle successive di tentare di descrivere questo modello attraverso gli strumenti dell'econometria, utilizzando un campione di dati fornito dal "Office for national statistics" del Regno Unito. Ovviamente, non essendo mai possibile creare un modello che descriva perfettamente la realtà, uno spazio importante è affidato alla verifica della validità e affidabilità del modello econometrico sviluppato. Per comprendere tutti questi argomenti, l'elaborato è suddiviso in cinque parti.

La teoria economica proposta nella prima parte è quella del "capitale umano" e del "modello dell'istruzione", che provano a spiegare come sia il livello di educazione a influire sui salari e come gli individui scelgono razionalmente il livello adatto per loro. Nella seconda parte si mostrano gli strumenti econometrici che saranno utilizzati nei capitoli successivi: il metodo dei minimi quadrati ordinari e il metodo delle variabili strumentali, in particolare i minimi quadrati a due stadi. La terza e la quarta parte riguardano la messa in pratica del modello: il terzo capitolo si occupa di illustrare le variabili inserite, la forma funzionale scelta, alcune limitazioni che possono scaturire dai dati, in particolare a causa dell'assenza di una variabile per l'abilità, e come si è tentato di superarle; il quarto mostra i risultati ottenuti e tenta di stabilirne la rilevanza e la veridicità, attraverso una serie di test statistici. Nell'ultima parte, infine, si approfondisce la criticità principale che emerge dal modello, ovvero il problema degli strumenti deboli.

La teoria del capitale umano

L'economia del lavoro è la "branca dell'economia politica che studia il funzionamento e le dinamiche del mercato del lavoro attraverso l'interazione tra lavoratori e imprese"(Treccani 2012). Tra le questioni che essa deve affrontare rientra la problematica della distribuzione dei redditi e della disparità dei salari percepiti dai diversi lavoratori. La teoria del capitale umano (si veda Borjas 2013) cerca di spiegare questo fenomeno in modo semplice. Ogni lavoratore porta con sé nel mercato del lavoro un insieme di abilità e capacità che si differenzia da quelle di ogni altro individuo. Queste caratteristiche, che possono essere innate o acquisite attraverso l'istruzione, formano il cosiddetto "capitale umano", che determina ciò che egli riesce a fare e come riesce a farlo: in altre parole, la sua produttività. Il reddito che ogni lavoratore percepisce, determinato dall'equilibrio tra domanda e offerta nel mercato del lavoro, è diretta conseguenza della produttività che egli può garantire: intuitivamente, per assicurarsi un lavoratore con un'alta produttività, il datore di lavoro è disposto a pagare relativamente di più. Questa prima parte si occupa perciò di stabilire come gli individui scelgano le capacità da acquisire e spendere nel mercato del lavoro, e come queste scelte incidano sui redditi che essi percepiranno lungo l'arco della propria vita lavorativa.

1 - Il modello dell'istruzione

Il modello dell'istruzione, proposto per primo da Mincer nel 1958 (si veda a proposito Borjas 2013) assume che gli individui considerano l'educazione scolastica unicamente come un mezzo per ottenere capitale umano: tutte le altre motivazioni che spingono un individuo a studiare vengono ignorate. L'istruzione è vista a tutti gli effetti come un investimento: ogni anno di istruzione aggiuntiva comporta dei costi immediati ma garantisce un incremento di capitale umano, il quale a sua volta permette di percepire redditi più elevati una volta iniziato a lavorare. Gli individui scelgono così il livello di istruzione che massimizza il valore attuale dei redditi lungo l'arco della vita lavorativa.

Disoccupazione	**2007**	**2014**
Licenza media	22,1%	48,1%
Diplomati	13,1%	30%
Laureati	9,5%	17,7%

(Elaborazione Almalaurea su dati Istat, 2015)

Reddito medio	**25-34**	**35-44**	**45-54**
Non laureati	22.543	25.917	27.479
Laureati	28.869	38.023	48.658

(Elaborazione Almalaurea su dati Istat, 2015)

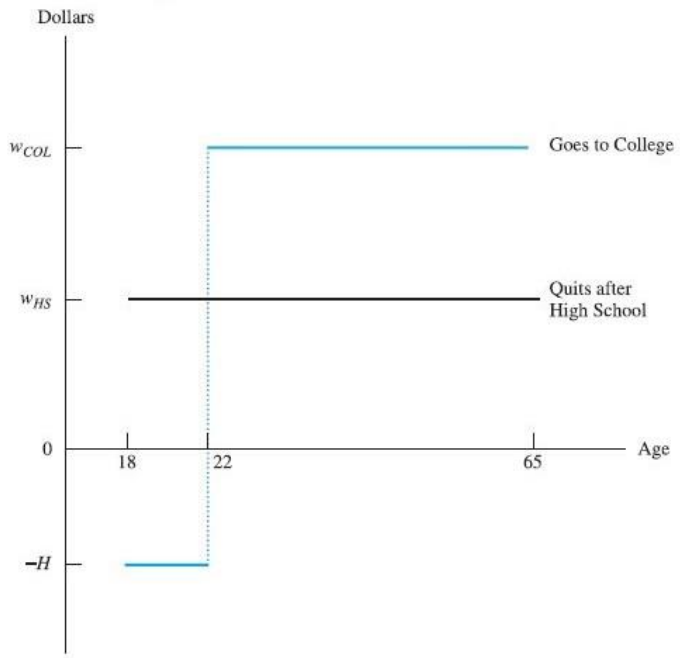
Come mostrano le tabelle sopra riportate, parrebbe evidente almeno in Italia la convenienza a proseguire gli studi fino all'ottenimento di una laurea, piuttosto che fermarsi al diploma o

addirittura prima. Non solo il reddito percepito è sensibilmente maggiore, ma anche il tasso di disoccupazione scende man mano che si prosegue con gli studi.

A un'analisi superficiale questi risultati possono sembrare inconciliabili con la teoria sopra esposta: se gli individui scelgono il livello di istruzione che massimizza il proprio reddito, come mai molti si fermano prima del raggiungimento della laurea? In realtà, la questione è molto più complessa.

Uno studente che decidesse di iscriversi all'università incorrerebbe in due tipi di costi: uno diretto, dato dalla somma delle tasse, libri, eventuale alloggio e ripetizioni; uno indiretto, il costo-opportunità derivante dal non percepire uno stipendio durante gli anni dell'università. Allo stesso tempo, tuttavia, una volta finiti gli studi percepirebbe un reddito più elevato (si veda *fig. 1*).

Fig. 1 – I potenziali flussi di reddito futuri di un neodiplomato



Fonte: Borjas 2013

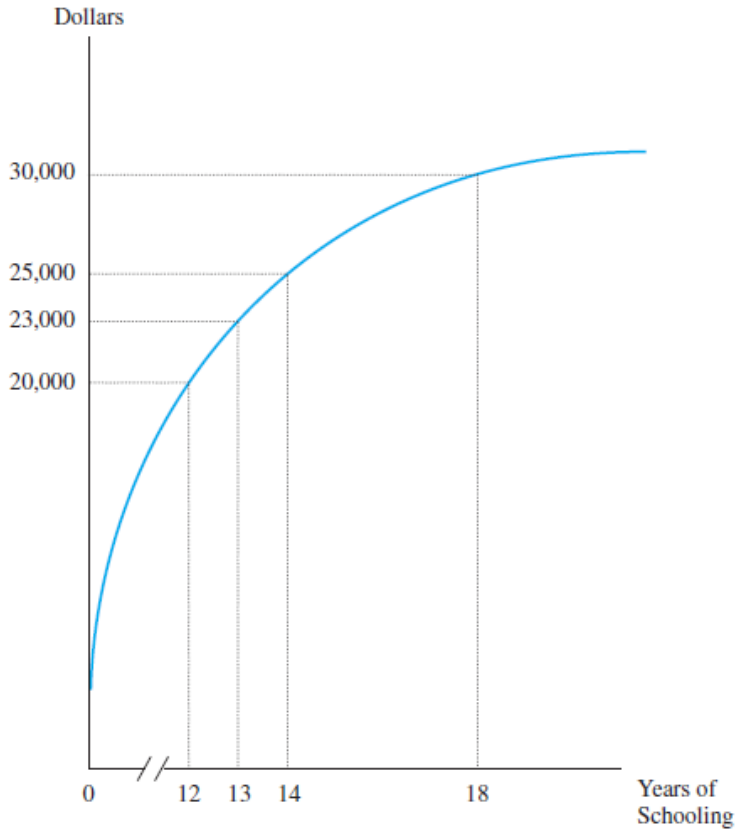
I valori attuali dei flussi di reddito derivanti dalle due scelte, assumendo per semplicità redditi costanti, sono dati dalle formule seguenti:

$$VA_{Scuola} = R_{Scuola} + \frac{R_{Scuola}}{1+r} + \frac{R_{Scuola}}{(1+r)^2} + \cdots + \frac{R_{Scuola}}{(1+r)^N}$$

$$VA_{Laurea} = -C - \frac{C}{1+r} - \frac{C}{(1+r)^2} + \frac{R_{Laurea}}{(1+r)^3} + \cdots + \frac{R_{Laurea}}{(1+r)^N}$$

Come si vede, quindi, il valore attuale netto derivante dalle diverse scelte di istruzione dipende da tre fattori: i costi sostenuti, il reddito futuro e, non ultimo, il tasso di sconto.

Fig. 2 – La relazione istruzione-reddito



Fonte: Borjas 2013

Il grafico sopra riportato confronta gli anni di istruzione con il reddito annuo corrispondente. Se assumiamo che l'unico costo dell'istruzione sia il costo-opportunità di non percepire uno stipendio, il grafico ha le seguenti proprietà (Borjas 2013):

- La pendenza è positiva: ciò significa che ogni anno aggiuntivo di istruzione comporta un incremento del reddito. Se ciò non fosse vero, nessuno studierebbe di più.

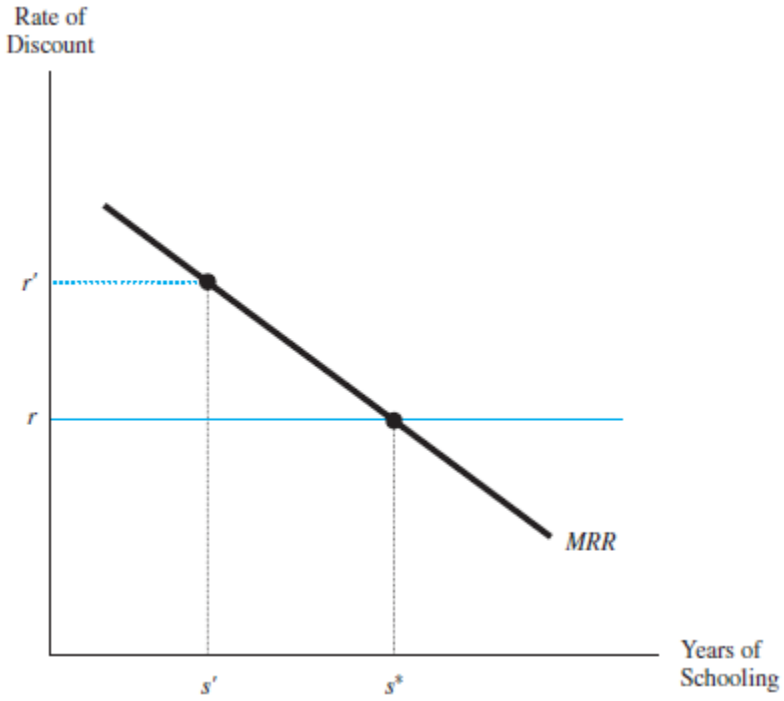
- È concavo: ciò significa che ogni anno aggiuntivo di istruzione incrementa il reddito in misura sempre minore. Anche questo è coerente con la teoria e con le conclusioni ottenute dalla letteratura (si veda ad esempio Psacharapoulos 1985).

- La pendenza è strettamente legata al rendimento marginale dell'istruzione, ovvero la variazione percentuale nei redditi data da un anno aggiuntivo di istruzione. Infatti, la pendenza della curva indica la variazione dei redditi in termini assoluti. Nel caso del 13esimo anno, essa è pari a (23.000 – 20.000)/(13 – 12) = 3.000. La variazione in percentuale è allora pari a 3.000/20.000 = 15%: questo è il rendimento marginale del 13esimo anno di istruzione. Esso coincide col rendimento di ogni dollaro speso in istruzione: per quell'anno infatti il costo sostenuto, a fronte di un aumento del reddito annuo di 3.000 dollari, è pari al reddito non

percepito, ovvero 20.000 dollari. Il rendimento dell'investimento è quindi pari, di nuovo, a 3.000/20.000 = 15%.

A partire da questa relazione se ne può calcolare una seconda (si veda *fig. 3)*, nella quale ad ogni anno di istruzione è associato il relativo rendimento marginale. Come si vede, il rendimento marginale è decrescente: questo deriva dalla concavità del grafico precedente. Questo grafico permette di stabilire, conoscendo il tasso di sconto di un individuo, il livello di istruzione ottimale: quello cioè in cui il rendimento marginale dell'istruzione è pari al tasso di sconto stesso (Borjas 2013). Un anno in più, o in meno, di istruzione rispetto al livello così stabilito comporterebbe una diminuzione del valore attuale dei redditi futuri. Come si desume dal grafico, più alto è il tasso di sconto minore è l'istruzione ottenuta. Un alto tasso di sconto significa minor peso dato ai redditi futuri rispetto a quelli presenti, con la conseguenza di preferire la rinuncia a maggiori guadagni futuri per ottenerne nel presente.

Fig. 3 – Il rendimento marginale dell'istruzione

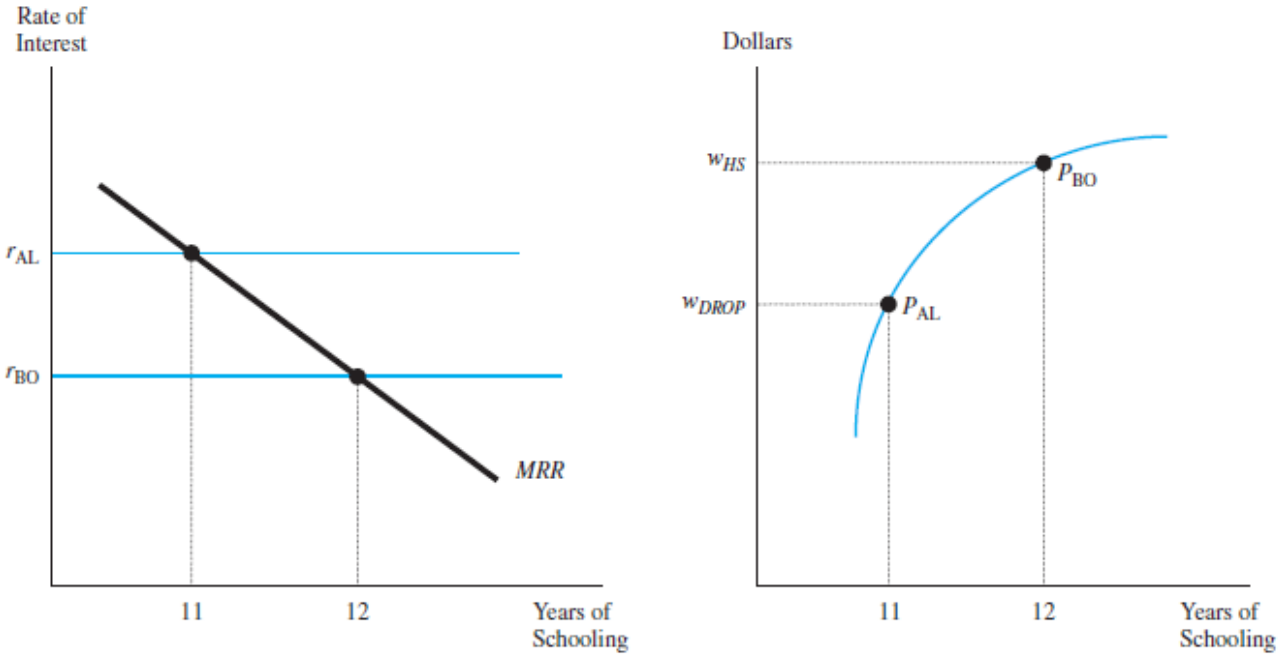


Fonte: Borjas 2013

Il modello dell'istruzione così descritto evidenzia i due fattori chiave che incidono nella scelta di quanta istruzione godere: il tasso di sconto e la curva del rendimento marginale. I paragrafi seguenti trattano le conseguenze pratiche nella determinazione del rendimento dell'istruzione dell'avere diversi tassi di sconto o curve del rendimento marginale.

1.1 Differenze nel tasso di sconto

Fig. 4 – Il rendimento dell'istruzione e la relazione istruzione-reddito per due individui con diverso tasso di sconto

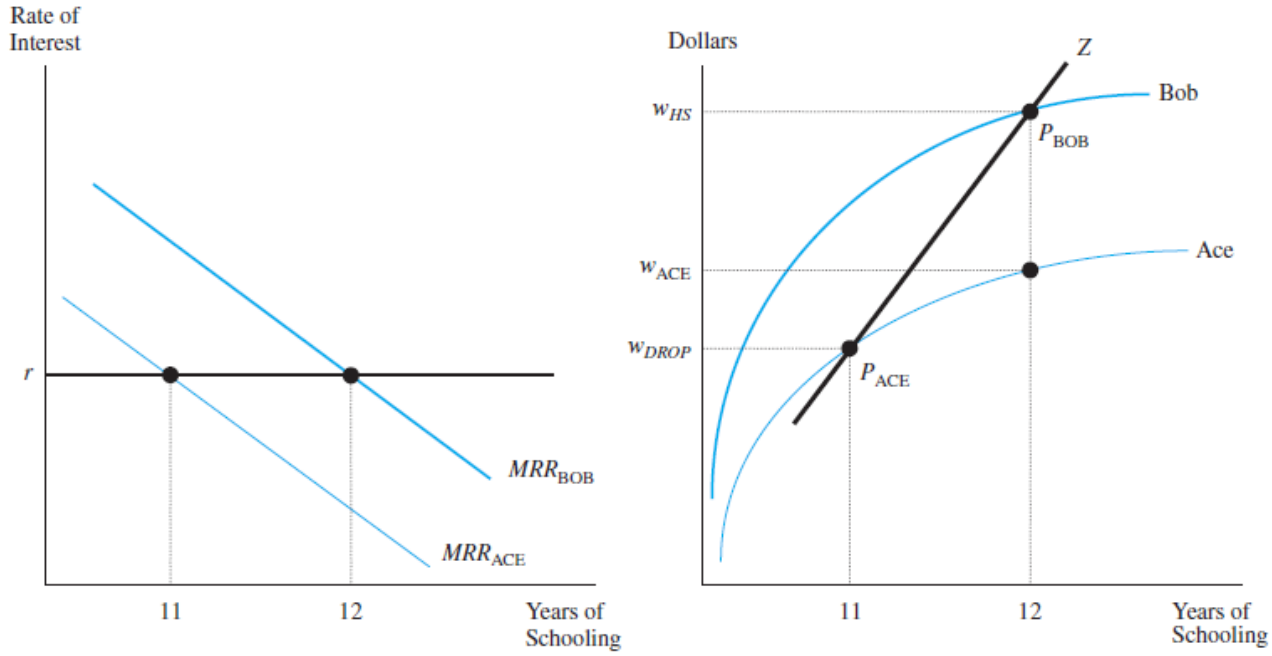


Fonte: Borjas 2013

Non è chiaro come si determini il tasso di sconto di un individuo, e da cosa esso dipenda (Borjas 2013). Assumendo però di conoscere che tra due individui a cambiare sia solo il tasso di sconto, significa che costoro fronteggiano la stessa curva di rendimento marginale. A causa del diverso tasso di sconto, tuttavia, la quantità di istruzione scelta è diversa: minore per l'individuo col tasso più alto, maggiore per l'altro (si veda il grafico a sinistra in *fig. 4)*. Se così fosse, il rendimento di un anno aggiuntivo di istruzione sarebbe immediatamente calcolabile conoscendo unicamente il reddito percepito e gli anni di istruzione dei due soggetti (si veda il grafico a destra). La condizione che a cambiare sia solo il tasso di sconto e che la curva del rendimento marginale dell'istruzione sia sempre la stessa, tuttavia, è estremamente limitante. Nel paragrafo seguente viene illustrato un motivo del perché la curva possa effettivamente essere diversa, e le conseguenze di questo fatto.

1.2 Differenze nella curva del rendimento marginale: il problema dell'abilità innata

Fig. 5 – Il rendimento dell'istruzione e la relazione istruzione-reddito per due individui con diversa curva dei rendimenti



Fonte: Borjas 2013

Le abilità possedute da un individuo non sono solamente ottenute attraverso l'istruzione: alcune sono "innate", presenti nel corredo genetico dell'individuo. Comunemente si ritiene che un più alto livello di abilità "sposti" la curva dei rendimenti verso destra, o detto in altro modo che ogni anno di istruzione aggiuntivo comporti un incremento di reddito maggiore, spingendo quindi l'individuo a studiare di più (Borjas 2013). Allo stesso modo un individuo con una bassa abilità trova presumibilmente più faticoso, e dunque più costoso, studiare, diminuendone ancora la convenienza. Tutto questo significa che, ad esempio, un individuo che abbia un basso livello di abilità innata e che quindi scelga razionalmente un basso livello di istruzione se fosse "costretto" a studiare quanto un individuo con un'alta abilità non arriverebbe a percepire lo stesso reddito. Al contrario, qualcuno ha suggerito (si veda a questo proposito Trostel et al. 2002) che l'istruzione porterebbe più benefici ai soggetti meno abili, in quanto quelli più bravi possiedono già alcune delle capacità trasmesse con l'istruzione. Altri ancora sostengono che un individuo portato al lavoro di intelletto, e che quindi abbia studiato maggiormente di un soggetto adatto ai lavori manuali, se svolgesse il lavoro del secondo individuo sarebbe meno bravo, e dunque verrebbe pagato meno. Ovunque risieda la verità tra queste teorie contrastanti, il risultato pratico che ne scaturisce è univoco: è impossibile determinare il rendimento marginale dell'istruzione basandosi solo sulla conoscenza dei redditi e degli anni di istruzione degli individui. Per capirlo, basta osservare il grafico (*fig. 5)* in cui sono state riportate le diverse curve di rendimento per due individui. Se prendessimo i

rispettivi redditi e anni di istruzione e provassimo a calcolare il rendimento marginale come in precedenza, otterremmo un risultato distorto. Nella seconda parte dell'elaborato saranno esposti i metodi statistici sviluppati per superare questo problema e cercare di produrre stime il più possibili corrette.

2 - La teoria dei segnali

Un'altra teoria parallela al modello dell'istruzione è la teoria dei segnali (Borjas 2013). L'ipotesi sottostante è che l'istruzione, contrariamente a quanto assunto dal modello dell'istruzione, non contribuisca ad aumentare in alcun modo la produttività degli individui, che è fissata. L'utilità dell'istruzione sarebbe quella di permettere di segnalare ai datori di lavoro il livello di produttività innato a ciascun lavoratore. Dal momento, infatti, che nel mercato del lavoro è presente un'asimmetria informativa tra i datori di lavoro e i lavoratori, sarebbe difficile per i primi separare i lavoratori con alta da quelli con bassa produttività, e il risultato sarebbe un livellamente dei salari e un'assegnazione alle mansioni del tutto casuale, con una forte perdita di efficienza. Per i lavoratori ad alta produttività, dunque, è conveniente segnalare ai datori di lavoro la propria abilità raggiungendo un alto grado di istruzione. Questo non succederebbe invece per gli individui meno abili, che troverebbero molto costoso studiare di più. Dal momento che i due modelli portano allo stesso risultato (redditi più alti per individui che studiano di più), non è facile determinare quale dei due sia quello corretto, o meglio quale dei due aspetti concorra di più nella determinazione del rendimento dell'istruzione. Tuttavia, è facile presumere che se l'educazione non contribuisse ad aumentare la produttività, ma solo a segnalarne il livello, sarebbero sorte aziende specializzate nel certificare l'abilità di un lavoratore in maniera meno costosa e impegnativa, soprattutto da un punto di vista temporale, del frequentare la scuola e l'università.

Gli strumenti econometrici

In questa sezione vengono esposti i metodi econometrici utilizzati nel seguito dell'elaborato. Innanzitutto viene illustrato il famoso metodo dei minimi quadrati ordinari (in inglese OLS: ordinary least squares) per poi passare alla spiegazione di metodi più complessi riguardanti il campo delle variabili strumentali, in particolare i minimi quadrati a due stadi (two stage least squares).

1 - Il metodo dei minimi quadrati

Il metodo dei minimi quadrati è lo strumento per eccellenza dell'econometria nel campo della regressione lineare (si veda ad esempio Gujarati & Porter 2010). Nel caso più semplice di un modello a due variabili, regressione lineare significa calcolo dei coefficienti della retta di miglior approssimazione. Il senso della dipendenza tra le due variabili deve essere suggerito da una sottostante teoria: nel caso di un modello sul rendimento dell'istruzione, è il reddito che "dipende" dagli anni di istruzione, e non viceversa. La retta di regressione risultante apparirebbe dunque così:

$$\hat{Y}_i = B_1 + B_2 X_i$$

Y= Reddito, X= Anni di istruzione.

Questa relazione deterministica, in cui il reddito è determinato esattamente dagli anni di istruzione, è ovviamente irrealistica. Un ruolo chiave nei modelli statistici è infatti ricoperto dal termine d'errore, ovvero la distanza verticale tra il valore "previsto" dalla retta di regressione e quello effettivamente osservato. Questo termine di errore rappresenta l'insieme degli altri fattori che concorrono a determinare il valore della Y ma non sono inclusi nel modello. La specificazione stocastica del modello precedente diventa come segue:

$$Y_i = B_1 + B_2 X_i + u_i$$

Y= Reddito, X= Anni di istruzione, u= Errore.

Il metodo dei minimi quadrati prescrive di scegliere B_1 e B_2 che minimizzino la somma dei quadrati degli errori. Matematicamente, ciò equivale a scrivere:

$$Min \sum u_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

$$= \sum (Y_i - B_i - B_2 X_i)^2$$

Attraverso calcoli algebrici si arriva a determinare il seguente valore dei coefficienti (Gujarati & Porter 2010):

$$B_1 = \bar{Y} - B_2\bar{X}$$

$$B_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \overline{X})^2}$$

Con $\bar{X}$ e $\bar{Y}$ a indicare il valore medio delle rispettive variabili.

Gli stessi procedimenti possono essere facilmente trasportati nei modelli a più variabili, dove la notazione si fa più complessa ma i concetti rimangono uguali.

In ambito pratico, tuttavia, la regressione viene applicata solo su un campione limitato di osservazioni. Convenzionalmente, la funzione di regressione del campione viene rappresentata così:

$$Y_i = b_1 + b_2X_i + e_i$$

Purtroppo, a meno che non vengano assunte determinate ipotesi, non è possibile stabilire come e se i coefficienti prodotti dalla regressione sul campione possano essere validamente estesi all'intera popolazione: questo è ciò che si propone di fare il modello classico di regressione lineare.

1.1 Il modello classico di regressione lineare

Le ipotesi sottostanti al modello classico di regressione lineare multipla sono le seguenti (Gujarati & Porter 2010):

1) Il modello è lineare nelle variabili
2) Le variabili esplicative sono incorrelate con il termine di errore *u*
3) Dato un certo valore delle X_i, il valore atteso di dell'errore *u* è 0. Cioè: $E(u|X_i) = 0$
4) La varianza di ogni *u* è costante, o omoschedastica: $var(u_i) = \sigma^2$
5) Non c'è correlazione tra due termini di errore, ovvero non c'è autocorrelazione: $cov(u_i, u_j) = 0\ i \neq j$
6) Il modello di regressione è correttamente specificato, vale a dire tutte le variabili che descrivono un determinato fenomeno sono incluse nel modello nella forma funzionale corretta
7) Non c'è esatta collinearità tra variabili indipendenti

Se queste ipotesi sono soddisfatte, b_1 e b_2 ed e_i sono stimatori corretti di B_1 e B_2 e u_i, i parametri della popolazione. Inoltre è possibile calcolare la varianza e l'errore standard degli stimatori, raccolti nella tabella.

	b_1	**b_2**	**e_i**
Var.	$\frac{\sum X_i^2}{n\sum X_i^2}\sigma^2$	$\frac{\sigma^2}{\sum X_i^2}$	$\frac{\sum e_1^2}{n-2}$
S.E.	$\sqrt{var(b_1)}$	$\sqrt{var(b_2)}$	$\sqrt{var(e_i)}$

Sotto queste ipotesi, il metodo OLS è particolarmente indicato grazie ad alcune proprietà molto positive, riassunte nel teorema di Gauss-Markov (Gujarati & Porter 2010):

1) b_1 e b_2 sono stimatori lineari, vale a dire funzioni lineari della variabile Y.
2) Sono corretti: $E(b_1)= B_1$ e $E(b_2)= B_2$.
3) $E(\hat{\sigma}^2) = \sigma^2$: lo stimatore della varianza dell'errore è corretto.
4) b_1 e b_2 sono stimatori efficienti. Questo significa che la loro varianza è minore di quella di qualsiasi altro stimatore di B_1 e B_2 rispettivamente.

Queste proprietà sono racchiuse nella sigla BLUE (Best Linear Unbiased Estimators).

Un'ultima ipotesi, fondamentale per condurre una serie di test statistici, è la seguente (Gujarati & Porter 2010):

8) Il termine di errore u_i segue la distribuzione normale con media 0 e varianza σ^2. Cioè:

$$u_i \sim N(0, \sigma^2)$$

Il fondamento logico sottostante a questa ipotesi deriva dal teorema centrale del limite, che stabilisce che in presenza di un grande numero di variabili casuali identicamente e indipendentemente distribuite la distribuzione della loro somma tende a essere normalmente distribuita mano a mano che il numero di queste variabili aumenta (si veda nuovamente Gujarati & Porter 2010). Dal momento che la natura del termine di errore è esattamente l'essere la somma di variabili casuali indipendenti, possiamo assumere senza particolari forzature che u_i sia distribuito secondo una normale.

Nel paragrafo successivo saranno trattate le conseguenze teoriche derivanti dalla violazione delle ipotesi sopra elencate.

1.2 Conseguenze delle violazioni delle ipotesi

Nel precedente paragrafo sono state elencate una serie di ipotesi necessarie per la valenza teorica del metodo OLS. In questo paragrafo saranno descritte, senza dimostrarle, le conseguenze che si verificano se vengono violate, e come e se è possibile contenerle.

1.2.1 Errori di specificazione

Questa situazione riguarda l'ipotesi della corretta specificazione del modello. Come si vedrà, essa riveste una particolare rilevanza per il modello analizzato nel seguito dell'elaborato. Come riportano Gujarati e Porter (2010), affinché un modello sia correttamente specificato occorre che varie condizioni siano soddisfatte. Di conseguenza, gli errori di specificazione possono essere di vario tipo.

Innanzitutto, è necessario che tutte le variabili rilevanti siano incluse nel modello. Se una variabile viene erroneamente omessa, le conseguenze possono essere più o meno gravi, a seconda se la variabile esclusa è correlata con una o più variabili. In caso positivo, gli stimatori dei coefficienti delle variabili correlate e dell'intercetta sono distorti e inconsistenti. In caso negativo, solo l'intercetta presenta uno stimatore distorto e inconsistente. In entrambi i casi, però, gli stimatori della varianza del termine di errore u_i e dei coefficienti sono distorti: questi ultimi, in particolare, presentano una distorsione positiva. Di conseguenza, gli intervalli di confidenza saranno più larghi, e tenderemmo ad accettare più frequentemente le ipotesi nulle (si veda il capitolo quarto).

Un secondo errore di specificazione del modello incorre quando vengono incluse variabili non necessarie. Quando ciò accade, gli stimatori dei coefficienti e della varianza dell'errore rimangono corretti. Tuttavia, le varianze degli stimatori dei coefficienti sono più grandi di quanto accadrebbe con il modello corretto: ciò significa di nuovo che gli intervalli di confidenza sono più larghi, anche se ancora accettabili. In altre parole, gli stimatori sono LUE, ma non BLUE: non sono efficienti.

Un errore facile da commettere riguarda la forma funzionale del modello. Nella quasi totalità dei casi, infatti, la teoria sottostante anche se è in grado di indicare le variabili da inserire e il segno corretto dei coefficienti non lo è per quanto riguarda la forma funzionale da adottare. Tuttavia, la scelta di una forma funzionale errata ha conseguenze gravi: i coefficienti stimati possono essere stime distorte dei veri coefficienti.

Infine, un errore da non dimenticare riguarda la stessa misurazione dei dati. Le conseguenze di errori di misurazione dei dati differiscono a seconda che essi riguardino la variabile dipendente o le variabili esplicative. Nel primo caso, l'unica conseguenza è che le varianze

stimate dei coefficienti sono più larghe: questo avviene perché l'errore nella variabile dipendente si aggiunge all'errore u_i. Se, al contrario, l'errore di misurazione riguarda le variabili indipendenti, gli stimatori sono distorti e inconsistenti.

Come si nota dalle discussioni precedenti, le conseguenze di avere uno o più errori di specificazione possono essere davvero serie. Fortunatamente, la teoria statistica mette a disposizione alcuni test per rilevare la presenza di alcuni di questi errori: alcuni verranno esposti e utilizzati nel modello sviluppato in questo elaborato. Alcuni di questi errori hanno un rimedio immediato: ad esempio, nel caso dell'inserimento di una variabile non rilevante, il rimedio ovvio è di escluderla dal modello. Tuttavia non è sempre così semplice. In caso di omissione di una variabile rilevante, infatti, se essa non è ottenibile in alcun modo, la soluzione diventa più complicata. Come si vedrà, in questo caso può diventare necessario ricorrere al metodo delle variabili strumentali.

1.2.2 Multicollinearità

Questa situazione si verifica quando due o più variabili indipendenti sono correlate linearmente. Il caso di perfetta multicollinearità, ovvero quando la correlazione è esatta, sebbene particolarmente grave perché non permette di stimare i coefficienti di ciascuna variabile, non si incontra facilmente nella realtà (Gujarati & Porter 2010). È invece più facile imbattersi nella situazione di imperfetta, ma alta, multicollinearità, ovvero quando le variabili indipendenti presentano un alto grado di correlazione. In questo caso le conseguenze sono meno gravi, anzi, gli stimatori rimangono BLUE, ma comunque non trascurabili. Infatti la varianza e gli errori standard degli stimatori, seppur efficienti, sono "grandi", con la conseguenza di produrre intervalli di confidenza più larghi. Questo può portare a pochi coefficienti significativi o addirittura al segno sbagliato degli stessi. (Gujarati & Porter 2010)

I rimedi più immediati di questo problema possono essere, a seconda del caso, escludere la variabile più altamente correlata dal modello, ottenere nuovi dati o un nuovo campione, ripensare la forma funzionale del modello o, infine, trasformare le variabili originarie (per una discussione più approfondita si veda sempre Gujarati & Porter 2010).

1.2.3 Eteroschedasticità

La condizione di omoschedasticità stabilisce che tutti i termini di errore u_i abbiano la stessa varianza σ^2. Se ciò non avviene, si incorre nella situazione di eteroschedasticità. Solitamente questo problema si verifica più nelle cross-section che nelle serie storiche. Questo avviene per via del cosiddetto "effetto scala": la variabile osservata in questo tipo di analisi ha solitamente valori molto variabili ed è quindi facile che l'ampiezza dell'errore aumenti mano a mano che

il valore delle variabili sottostanti aumenta. Proprio per questo motivo, la varianza tende ad aumentare o diminuire al variare delle X. (Gujarati & Porter 2010)

Le conseguenze dell'eteroschedasticità sono che gli stimatori OLS sono ancora lineari e corretti, ma non hanno più varianza minima. Inoltre, le stime delle varianze degli stimatori sono generalmente distorte: questo accade dal momento che lo stimatore della varianza dell'errore non è più corretto. Come conseguenza di ciò, gli intervalli di confidenza non sono più affidabili. (Gujarati & Porter 2010)

Una soluzione al problema sviluppata da White è di calcolare gli errori standard degli stimatori attraverso una formula differente, che tenga conto dell'eteroschedasticità. Gli errori standard così prodotti vengono chiamati "errori standard robusti" e permettono di sviluppare i test statistici necessari in modo affidabile con qualunque tipo di eteroschedasticità, compreso il caso particolare di omoschedasticità. (Gujarati & Porter 2010)

1.2.4 Autocorrelazione

In una regressione si ha autocorrelazione quando i termini di errore sono correlati in modo seriale. Al contrario dell'eteroschedasticità, l'autocorrelazione si presenta più frequentemente nelle serie storiche. Matematicamente, si può esprimere questa situazione nel modo seguente:

$$E(u_i, u_j) \neq 0 \quad i \neq j$$

L'autocorrelazione può essere sia positiva che negativa, e può accadere per svariati motivi, sia legati al fenomeno in sé (concetto di inerzia) che a errori di specificazione del modello. Le conseguenze della correlazione seriale tra termini di errore sono che gli stimatori rimangono lineari e corretti, ma non efficienti: vale a dire che non hanno varianza minima. Inoltre, le formule per calcolare la varianza degli stimatori e del termine di errore sono distorte: di conseguenza gli intervalli di confidenza calcolati sulla loro base non sono affidabili, così come l'R^2. (Gujarati & Porter 2010)

1.2.5 Correlazione tra variabili esplicative e termine di errore

I motivi per cui una variabile esplicativa possa essere correlata al termine d'errore sono diversi, alcuni di essi sono già stati esposti in precedenza. Innanzitutto, è possibile che questo accada quando una variabile rilevante viene omessa dal modello (Wooldridge 2003). Consideriamo il modello ristretto:

$$y_i = \beta x_i + u_i$$

Mentre quello corretto sarebbe:

$$y_i = \propto_1 x_i + \propto_2 w_i + v_i$$

Con qualche calcolo si può mostrare che l'errore del primo modello è dato da (si veda Wooldridge 2003):

$$u_i = \left[w - x\left(\frac{Cov(x,w)}{Var(x)}\right)\right] \propto_2 + v_i$$

Come si vede, esso è correlato con la variabile *x*, a meno che *x* e *w* non siano incorrelate tra loro (*Cov(x,w)=0)* o che a_2 non sia pari a 0, cosa che però avviene solo se la variabile w è irrilevante, e quindi giustamente omessa. Come già detto, in questi casi β è uno stimatore distorto del vero parametro della popolazione.

Come detto, le cause di una correlazione tra variabile esplicative e termine di errore non si fermano qui. Altre possono essere la presenza di equazioni simultanee, in cui cioè la Y non è unilateralmente determinata dalla X ma concorre a sua volta a determinare la stessa X, o errori nella determinazione della variabile esplicativa (per una discussione più approfondita, si rimanda al già citato Wooldridge 2003). Nel modello sviluppato nelle parti successive dell'elaborato sarà importante però proprio la prima situazione considerata. Come infatti spiegato nella prima parte, non è possibile determinare in modo immediato il rendimento marginale dell'istruzione se non si è in possesso di una misura della variabile "abilità innata". In questo caso infatti regredendo semplicemente il reddito sugli anni di istruzione non si tiene conto dell'influenza dell'abilità nella determinazione del rendimento dell'istruzione, e di conseguenza nella scelta degli anni di istruzione. Così facendo si rientrerebbe esattamente nel caso sopra illustrato, con stimatori distorti e inconsistenti. Per risolvere il problema, due sono le soluzioni possibili. La prima è di trovare una "variabile proxy", che funga cioè da sostituta della variabile mancante. Perché la variabile così trovata sia efficace, occorre che presenti un alto grado di correlazione con quella omessa. Una seconda soluzione consiste invece nell'utilizzare un metodo diverso da quello dei minimi quadrati ordinari, che sia in grado di produrre stime consistenti. Questo metodo verrà spiegato nel paragrafo successivo.

2 - Il metodo delle variabili strumentali

Questo metodo permette di calcolare consistentemente il coefficiente della variabile correlata con l'errore. Ipotizziamo di avere il seguente semplice modello:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

tale che *x* sia correlata con l'errore *u* ($Cov(x,u) \neq 0$), a causa dell'omissione di una variabile rilevante.

Il metodo delle variabili strumentali richiede di individuare una variabile, chiamata strumento (z), che soddisfi le seguenti condizioni (Wooldridge 2003):

$$Cov(z, u) = 0$$

$$Cov(z, x) \neq 0$$

La prima condizione si traduce con l'espressione "variabile esogena", e implica che la variabile strumentale (IV) *s* sia incorrelata con la variabile omessa e che quindi non abbia effetto su *y* (dopo aver controllato per *x).* La seconda indica semplicemente che *s* deve essere correlata in qualche modo a *x*. Ciò significa che nel modello in esame in questo elaborato lo strumento deve essere correlato con gli anni di istruzione ma non con l'abilità innata. La seconda condizione può essere verificata facilmente stimando la seguente equazione:

$$x_i = \pi_0 + \pi_1 z_1 + v_i$$

Se π_1 risulta statisticamente diverso da 0, la condizione è verificata. A questo punto si può mostrare come lo stimatore IV di β_1 sia:

$$\hat{\beta}_1 = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})}$$

Quello di β_0 rimane lo stesso ($\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$) ma usando lo stimatore IV di $\hat{\beta}_1$ invece che l'OLS. È importante sottolineare che gli stimatori IV sono corretti solo asintoticamente, ovvero sono consistenti. Ciò significa che per produrre stime non distorte è necessario lavorare con un campione numeroso. (Wooldridge 2003)

Una precisazione importante da fare prima di procedere oltre con il metodo delle variabili strumentali è che la varianza degli stimatori IV è diversa da quella ottenuta con il metodo OLS, e in particolare è sempre maggiore. Guardiamo perché: l'ipotesi iniziale è quella di omoschedasticità, ovvero $Var(u|z)=\sigma^2$. Si dimostra allora che la varianza di β_1 è (si veda Wooldridge 2003 per un'analisi più approfondita):

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{n\sigma_x^2 \rho^2_{x,z}}$$

con $\rho^2_{x,z}$ uguale al quadrato del coefficiente di correlazione tra x e z. Lo stimatore corrispettivo è:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{TSS_x R^2_{x,z}}$$

Ricordando che la varianza del coefficiente β_1 con gli OLS è data da $\frac{\hat{\sigma}^2}{TSS_x}$, si vede che l'unica differenza è data dal termine $R^2_{x,z}$ al denominatore: dal momento che $0 \leq R^2 \leq 1$, è facile capire che se *x* e *z* non sono perfettamente correlate (in questo caso i due metodi coinciderebbero) la varianza IV è sempre maggiore di quella OLS. Minore è l'$R^2_{x,z}$ maggiore è la varianza IV. Il problema di avere una bassa correlazione tra la variabile endogena e il suo strumento non si risolve, tuttavia, solo in una grande varianza. Infatti, se accade che *z* e *u* siano anche solo poco correlati, lo stimatore IV può presentare una grande distorsione anche asintoticamente, ovvero essere inconsistente. Per capire il perché, si mostra il valore dello stimatore $\hat{\beta}_1$ quando *z* e *u* sono correlati (si rimanda di nuvo a Wooldridge 2003 per la derivazione):

$$plim\hat{\beta}_{1,IV} = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

La parte interessante risiede nelle correlazioni. Si capisce come anche se *Corr(z,u)* è piccola, l'inconsistenza dello stimatore IV può essere molto grande se anche *Corr(z,x)* è modesta. Ciò significa che potrebbe essere meglio utilizzare il metodo OLS. Questo problema, chiamato degli strumenti deboli, verrà ripreso e approfondito più avanti.

2.1 Metodo delle variabili strumentali in una regressione multipla

Il metodo precedente può essere esteso agevolmente alle regressioni multiple. Consideriamo il seguente modello, che prende il nome di *equazione strutturale*:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

Di nuovo abbiamo una variabile omessa correlata con y_2, tale che $Corr(y_2,u) \neq 0$, $Corr(y_1,u) \neq 0$. La variabile z_1, invece, è esogena. Consideriamo inoltre la variabile esogena z_2, esclusa dall'equazione strutturale ma correlata con y_2. Abbiamo quindi le seguenti condizioni:

$$\begin{cases} E(u_1) = 0 \\ Cov(z_1, u_1) = 0 \\ Cov(z_2, u_1) = 0 \end{cases} \rightarrow \begin{cases} E(u_1) = 0 \\ E(z_1 \cdot u_1) = 0 \\ E(z_2 \cdot u_1) = 0 \end{cases}$$

$\hat{\beta}_0, \hat{\beta}_1, e\ \hat{\beta}_2$ si ottengono risolvendo le controparti del campione:

$$\sum (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum z_{i1} \cdot (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum z_{i2} \cdot (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

Come detto, z_2 deve essere correlata con y_2, ma in questo caso si parla di correlazione parziale, ovvero tolto l'effetto di z_1. Di nuovo, il modo più immediato per vederlo è scrivere quella che si chiama *equazione del primo stadio*:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$$

La condizione di identificazione è $\pi_2 \neq 0$, ovvero y_2 correlata con z_2 anche dopo aver corretto per z_1. Anche avendo più variabili esogene comprese nell'equazione strutturale il metodo è immediatamente applicabile. Le condizioni rimangono le stesse: tutte le variabili esogene comprese nell'equazione strutturale devono essere incorrelate con l'errore, mentre nell'equazione del primo stadio è sufficiente che il coefficiente dello strumento escluso sia correlato con la variabile endogena, per gli altri non è necessario. (Wooldridge 2003)

2.2 I minimi quadrati a due stadi

Finora abbiamo considerato il caso in cui ci fosse solo una variabile esogena esclusa dall'equazione strutturale e utilizzabile come strumento. Il metodo delle variabili strumentali può però essere utilizzato anche nel caso in cui ci fossero più di una variabile esogena potenzialmente adatta a fungere da strumento (si vedano a riguardo Wooldridge 2003 e Cappuccio & Orsi 1995). In questa situazione, il metodo prende il nome di *minimi quadrati a due stadi* (2SLS dall'inglese *two stage least squares*).

Ancora una volta consideriamo un'equazione strutturale del tipo:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

In cui, a causa di una variabile omessa, la variabile y_2 è endogena per cui $Corr(y_2,u_1) \neq 0$, $Corr(y_1,u_1) \neq 0$. Questa volta però abbiamo abbiamo due variabili esogene escluse, z_2 e z_3, correlate con y_2. Ciò significa che ogni combinazione lineare tra z_1, z_2 e z_3 è una valida IV. La scelta migliore è quindi quella di scegliere la combinazione lineare maggiormente correlata con y_2: questa risulta essere proprio l'equazione del primo stadio

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$$

Dove $E(v_2) = 0$, $Cov(z_1,v_2) = 0$, $Cov(z_2,v_2) = 0$, $Cov(z_3,v_2) = 0$.

La migliore variabile strumentale per y_2 è quindi $y^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$. La condizione di identificazione richiede che $\pi_2 \neq 0$ o $\pi_3 \neq 0$, ovvero che almeno una delle due variabili strumentali escluse dall'equazione strutturale sia correlata con la variabile endogena

y_2. Si può interpretare l'equazione del primo stadio come un modo per "dividere" la variabile y_2 in due parti: y^*, incorrelata con u_{1i}, e v_2, che invece è correlata. Avendo z_1, z_2 e z_3 possiamo quindi calcolare y^*, posto di conoscere i veri π, e utilizzarla nell'equazione strutturale. Poiché però nella pratica non conosciamo i veri π dobbiamo stimare questa equazione attraverso il campione con OLS:

$$\hat{y}_{i2} = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

Una volta ottenuto questa nuova variabile, le equazioni attraverso le quali è possibile stimare β_0 β_1 β_2 sono:

$$\sum \left(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}\right) = 0$$

$$\sum z_{i1} \cdot \left(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}\right) = 0$$

$$\sum \hat{y}_{i2} \cdot \left(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}\right) = 0$$

Il modello econometrico

I dati utilizzati nel modello successivamente descritto provengono dal dataset del primo trimestre (gennaio – marzo) 2016 della "Labour force survey (LFS)", un'indagine condotta tra le famiglie residenti nel Regno Unito da parte del "Office for national statistics" (Office for national statistics 2016). Il suo scopo è di "fornire indicazioni sul mercato del lavoro britannico che possano essere usate per sviluppare e valutare le politiche del mercato del lavoro". La prima indagine della serie fu condotta nel 1973, dal 1992 è sviluppata su base trimestrale. Il campione indagato consta in 60000 famiglie. Nel trimestre di riferimento, è formato da 90787 individui, che si riducono a 9518 dopo averlo pulito dalle informazioni mancanti.

La forma funzionale scelta, coerentemente con l'ampia letteratura disponibile sul tema (tra gli altri, ad esempio, Card 2001; Leigh & Ryan 2008; Trostel et al. 2002), segue il modello di Mincer. Si tratta di una funzione semilogaritmica, con la variabile indipendente sotto forma di logaritmo naturale e le variabili esplicative che seguono una forma lineare o quadratica. I coefficienti di questa forma funzionale particolare esprimono la variazione relativa nella variabile dipendente dovuta a una variazione assoluta nella variabile esplicativa. In altre parole, se moltiplicati per cento, essi indicano la variazione percentuale del reddito dovuta a una variazione unitaria della variabile di riferimento. Le variabili esplicative usate sono di due tipi, quantitative o qualitative. Le variabili qualitative sono state inserite attraverso l'uso di una variabile dummy, che assume esclusivamente valore di 1 o 0. Nel caso di variabili qualitative con più di due categorie, è stata creata una dummy per ogni categoria meno una (per non incorrere nella "trappola delle dummy" (Gujarati & Porter 2010)).

Anche se la forma funzionale scelta è in coerenza, come detto, con quanto prodotto dalla letteratura, bisogna tenere in considerazione che essa non ricalca con precisione quanto teorizzato nel modello dell'istruzione riguardo al rendimento marginale dell'istruzione. Esso viene infatti assunto come decrescente, ma il modello semilogaritmico adottato produce come risultato un rendimento marginale costante in termini relativi, e di conseguenza addirittura crescente in termini assoluti. Questa semplificazione, tuttavia, non dovrebbe creare grossi problemi e ha il vantaggio di rispondere all'esigenza di ottenere una stima unica, che rappresenti in un certo modo la media dei rendimenti marginali per ogni tipo di individuo e per ogni anno di istruzione.

1 - Le variabili inserite

Il modello sviluppato segue la seguente equazione:

$$LNHOURPAY = \beta_0 + \beta_1 YEDUC + \beta_2 POTEXP + \beta_3 POTEXP2 + \beta_4 PTIME + \beta_5 PUBLIC + \beta_6 FEMALE + \beta_7 TRAINING + \beta_8 LONDON + \beta_9 SCOTLAND + \beta_{10} SOUTH + \beta_{11} COHABITANT + \beta_{12} BLACK + \beta_{13} ASIAN + \beta_{14} INDIAN$$

Le diverse variabili hanno il seguente significato:

LNHOURPAY: Il logaritmo naturale del salario orario

YEDUC: Gli anni di istruzione conseguiti

POTEXP: L'esperienza lavorativa (potenziale), calcolata sottraendo all'età attuale l'età in cui si è lasciata l'istruzione

POTEXP2: Il quadrato di POTEXP

PTIME: variabile dummy che assume valore di 1 se l'individuo ha un lavoro part-time, 0 per un lavoro full-time

PUBLIC: variabile dummy che assume valore di 1 se l'individuo lavora nel settore pubblico, 0 se lavora nel settore privato

FEMALE: variabile dummy che assume valore di 1 se l'individuo è una donna, 0 se è un uomo

TRAINING: variabile dummy che assume valore di 1 se l'individuo ha svolto durante la vita lavorativa un periodo di training o di formazione, 0 altrimenti

LONDON, SCOTLAND, SOUTH: variabili dummy che assumono valore di 1 se l'individuo risiede, rispettivamente, a Londra, in Scozia o nel sud dell'Inghilterra, 0 altrimenti

COHABITANT: variabile dummy che assume valore di 1 se l'individuo convive con un partner, 0 altrimenti

BLACK, ASIAN, INDIAN: variabili dummy che assumono valore di 1 se l'individuo dichiara di appartenere, rispettivamente, all'etnia nera, asiatica o indiana, 0 altrimenti

Tutte queste variabili sono state calcolate attraverso l'uso e la combinazione di una o più variabili del database, ovvero HOURPAY, FTPT, EDAGE, AGE, PUBLICR, SEX, ED13WK, URESMC, MARSTA, ETHUKEUL, RELIG11, NATOX7.

Tra queste variabili, la più importante è ovviamente YEDUC: la stima corretta del suo coefficiente è l'obiettivo primario del presente elaborato. Le altre variabili sono state inserite per "pulire" l'effetto dell'istruzione da altre variabili che presumibilmente, sulla base della

teoria economica o della letteratura, influenzano il reddito. L'esperienza lavorativa compare sia linearmente che in modo quadratico. Il motivo di questa scelta risiede nei rendimenti marginali attribuiti dalla teoria economica all'esperienza lavorativa, che sono ipotizzati decrescenti. Di conseguenza, il coefficiente previsto di POTEXP sarà positivo e quello di POTEXP2 negativo e di valore assoluto minore.

1.1 Le variabili mancanti

Come sappiamo, questo modello è inficiato dall'omissione di una variabile molto importante e correlata con l'istruzione, che possiamo chiamare "abilità innata". La letteratura ha fornito alcuni esempi di variabili utilizzate come proxy per l'abilità (si veda Borjas 2013). Tra queste compaiono ad esempio una misura del QI o di altri test di intelligenza, o i voti presi a scuola. Sono stati sollevati dubbi tuttavia sulla bontà di queste variabili come proxy. Anche per questa ragione, in questo elaborato verrà utilizzato il metodo delle variabili strumentali per produrre stime consistenti del rendimento dell'istruzione.

Oltre al problema della mancanza della variabile abilità, non è stato possibile inserire altre variabili che hanno presumibilmente un effetto sul reddito, a causa della loro assenza nel dataset. Tra queste, si citano una variabile per gli individui che abbiano vissuto un periodo di disoccupazione, una variabile per gli immigrati non madrelingua inglesi, una per chi ha o ha avuto in passato problemi di salute, una per l'appartenenza o meno al sindacato. Gli effetti di queste, e delle altre, variabili mancanti sono di conseguenza inclusi nel termine di errore. Tra queste variabili e l'abilità, tuttavia, c'è un'importante differenza: mentre la prima è molto probabilmente correlata con l'istruzione, con tutte le conseguenze in precedenza esposte, queste ultime non presentano ragioni apparenti per esserlo con nessuna delle variabili inserite.

2 - Gli strumenti per l'abilità

La letteratura propone alcuni esempi interessanti di variabili utilizzati come strumenti per l'istruzione. Di nuovo, una variabile per poter essere usata come strumento deve essere correlata con quella che si decide di strumentare, e incorrelata con la variabile omessa, e dunque con l'errore. Tra gli altri, Trostel, Walker e Woolley (2002) usano l'istruzione del partner o dei genitori, mentre Leigh e Ryan (2008) tentano di sfruttare le differenze negli anni di istruzione dovute al mese di nascita o a modifiche del numero legale di anni di istruzione minima. Taubman (1976) e Ashenfelter e Kruegel (1994), infine, utilizzano coppie di gemelli con anni di istruzione diversi.

In questo elaborato si è deciso di provare ad utilizzare come variabile strumentale il credo religioso. Ricordiamo le due condizioni necessarie perché una variabile possa essere usata

come strumento: correlazione con la variabile endogena, e incorrelazione con la variabile esclusa. Mentre per la validità della seconda ipotesi non sembrano esserci particolari dubbi, la seconda appare meno sicura. Nelle parti successive si tenterà di capire se questa condizione regga o meno, e le eventuali conseguenze. Essendo la religione una variabile qualitativa, è necessario utilizzare le variabili dummy. Sono state così definite le seguenti variabili:

MUSLIM: variabile che assume il valore di 1 se l'individuo si professa di religione musulmana, 0 altrimenti

HINDU: variabile che assume il valore di 1 se l'individuo si professa di religione induista, 0 altrimenti

BUDHIST: variabile che assume il valore di 1 se l'individuo si professa di religione buddista, 0 altrimenti

OTHER: variabile che assume il valore di 1 se l'individuo non si professa di nessuna delle religioni precedenti, né cristiano, né non credente, 0 altrimenti.

Nella parte seguente verranno esposti e interpretati i risultati della regressione ottenuta secondo il modello qui proposto e verranno formulati alcuni test per verificare le ipotesi sopra indicate.

Risultati e test

In questa parte vengono presentati in due paragrafi diversi i risultati ottenuti: il primo riguarda la semplice regressione con il metodo OLS, il secondo il metodo delle variabili strumentali che comprende i minimi quadrati a due stadi e il limited information maximum likelihood.

1 - Il metodo dei minimi quadrati

Nella tabella sottostante sono riportati i risultati ottenuti regredendo i dati attraverso il metodo OLS. Prima di passare alla spiegazione del significato delle altre colonne si cercherà di capire se i coefficienti stimati sono coerenti con le aspettative iniziali.

LNHOURPAY	Coef.	Std. Err.	t	P>|t|	95% Conf. Interval	
Constant	1.350416	.0338569	39.89	0.000	1.284049	1.416783
YEDUC	.0508385	.0018815	27.02	0.000	.0471504	.0545266
POTEXP	.0312425	.001594	19.60	0.000	.0281179	.0343671
POTEXP2	-.0004849	.0000307	-15.78	0.000	-.0005451	-.0004246
PTIME	-.2807831	.0128128	-21.91	0.000	-.3058989	-.2556672
PUBLIC	.1095704	.0122189	8.97	0.000	.0856187	.1335221
FEMALE	-.1310353	.0115509	-11.34	0.000	-.1536775	-.1083932
TRAINING	.0710621	.0119822	5.93	0.000	.0475744	.0945497
LONDON	.2588164	.0190248	13.60	0.000	.2215237	.2961092
SCOTLAND	.0618323	.0203671	3.04	0.002	.0219085	.1017561
SOUTH	.0955729	.0120843	7.91	0.000	.071885	.1192608
COHABITANT	.1701433	.0113759	14.96	0.000	.1478442	.1924424
BLACK	-.2215648	.035925	-6.17	0.000	-.2919856	-.1511441
ASIAN	-.239672	.0331589	-7.23	0.000	-.3046705	-.1746734
INDIAN	-.0997315	.0343847	-2.90	0.004	-.167133	-.0323301

R-squared	0.2442
Adj R-squared	0.2431
F(14, 9503)	219.34
Prob > F	0.0000

Il valore dell'intercetta ("Constant") non ha un'interpretazione immediata: rappresenta il logaritmo naturale del salario orario medio di un individuo con il valore di tutte le variabili uguale a zero. Con un rapido calcolo si trova che il salario orario medio di questi individui è pari a 3.86 sterline. Gli altri coefficienti sono più interessanti. Quello dell'istruzione è pari a circa 0.051: ciò significa che ogni anno aggiuntivo di istruzione aumenta il salario orario del 5.1%. Questo risultato è in linea con quanto emerso da altri studi simili (si veda Borjas 2013). Gli altri coefficienti hanno una interpretazione simile, in particolare sono interessanti quelli di POTEXP e POTEXP2, pari rispettivamente a circa 0.0312 e -0.0005. Questi valori confermano perciò le assunzioni fatte a priori sul rendimento marginale decrescente dell'esperienza. Anche per le altre variabili è importante guardare prima di tutto al segno, per capire se esso è coerente con le aspettative, se c'erano. Il segno negativo associato a FEMALE, BLACK, ASIAN, INDIAN è quello che si supponeva sulla base di ricerche

effettuate sul problema della discriminazione (si veda ad esempio il già citato Borjas 2013). Anche il coefficiente negativo di PTIME appare corretto, così come quelli positivi di TRAINING, LONDON E SOUTH.

Nella seconda tabella viene presentato il valore di due voci simili: "R-squared" e "Adjusted R-Squared". L'R-squared, o coefficiente di determinazione, ha valore che varia da 0 a 1, e indica quanta parte della variazione totale in Y (il logaritmo del salario orario) è spiegata dal modello. Molto velocemente, chiamando TSS (total sum of squares) la somma dei quadrati delle deviazioni degli Y dalla media, ESS (estimated sum of squares) la somma dei quadrati delle deviazioni degli $\hat{Y}$ (la stima di Y prodotta dal modello) dalla media e RSS (residuals sum of squares) la somma dei quadrati degli errori, R^2 è dato da:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Da queste formule si capisce che un valore di R^2 pari a 1 significa che gli Y seguono esattamente la regressione, ovvero che i residui sono sempre pari a 0. Viceversa, un valore pari a 0 significa che il modello non spiega in alcun modo la variazione degli Y. Ovviamente, nelle applicazioni pratiche tali valori esatti non si trovano, ma si possono avere R^2 che variano in questo intervallo: più vicini sono a 1, "migliore" è il modello. Nel caso presente, il valore di R^2 del 24.4% è relativamente basso: ciò può essere un indicatore del fatto che manchino alcune variabili rilevanti. L'adjusted R^2 viene calcolato, sulla base del R^2, per permettere il confronto di due modelli con più o meno variabili esplicative. Per permettere questo confronto è necessario infatti correggere il coefficiente dal fatto che esso aumenta sempre con l'aggiunta di una variabile esplicativa. I due valori in questo caso non differiscono di molto. (Gujarati & Porter 2010)

Ora è necessario ricordare che i valori dei coefficienti ottenuti sono semplici stime basate su un campione, non è detto che coincidano con i valori veri della popolazione. Fortunatamente è possibile stabilire, sulla base delle ipotesi formulate alla base del modello, quanto "buone" siano queste stime, e soprattutto se il loro valore è significativo, cioè "statisticamente" diverso da zero. Inoltre, è possibile che alcune delle ipotesi non siano rispettate: nei prossimi paragrafi si tenterà allora di testare queste ipotesi attraverso gli strumenti che la statistica mette a disposizione.

1.1 La significatività delle variabili: il test T di Student

Il test T ha lo scopo di verificare se il valore di un parametro si discosta o meno da un valore preso come riferimento. In questo caso, si cerca di capire se i valori dei parametri della regressione sono diversi da 0. Formalmente, si esprime in questo modo.

Sotto le ipotesi:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

La variabile:

$$t = \frac{b_i}{se(b_{i)}}$$

segue la distribuzione T di Student con gradi di libertà pari al numero delle osservazioni (9518 nel presente caso) meno il numero dei parametri della regressione, compresa l'intercetta (15) (Gujarati & Porter 2010). I valori della statistica sono elencati nella prima tabella per ciascun parametro. Alla destra, nella colonna P>|t|, è indicata la probabilità associata a questo valore nella distribuzione, ovvero la probabilità di ottenere quel valore del parametro supposta vera l'ipotesi nulla. Comunemente, una probabilità inferiore al 5% è considerata adeguata per poter affermare che il coefficiente è significativamente diverso da 0. Le ultime due colonne derivano sempre da questo test, e rappresentano il valore inferiore e superiore dell'intervallo di confidenza al 95% di confidenza. In parole povere, rappresentano l'intervallo entro al quale, in 95 volte su 100, è compreso il valore della stima del parametro prodotta da campioni diversi della popolazione. Nel modello in questione, tutte le variabili sono significative.

1.2 La significatività congiunta: il test F

Le ultime due righe della seconda tabella sono riferite al cosiddetto test F, che si propone di verificare che i parametri siano congiuntamente, e non singolarmente, diversi da 0, ovvero che il coefficiente di determinazione sia diverso da 0. Sotto le ipotesi:

$$H_0: R^2 = 0$$

$$H_1: R^2 \neq 0$$

La variabile:

$$F = \frac{ESS/d.f.}{RSS/d.f.}$$

segue la distribuzione F con al numeratore (k-1) = 14 e al denominatore (n-k) = 9503 gradi di libertà (Gujarati & Porter 2010). Come col test T, si può rifiutare l'ipotesi nulla se il valore p è inferiore al 5%. Nel presente caso, esso è inferiore allo 0,01%.

1.3 Eteroschedasticità: il test di Breusch-Pagan e il test di White

Come detto, nelle cross-section è probabile che la condizione che la varianza dell'errore non dipenda dalle variabili esplicative (omoschedasticità) non sia valida. Per questo, è opportuno testare questa ipotesi. Sono ormai molti i test sviluppati per questo scopo: qui vengono utilizzati il test di Breusch-Pagan e il test di White.

Consideriamo una normale regressione lineare, di cui assumiamo che tutte le condizioni siano verificate tranne quella di omoschedasticità che si vuole testare. Ora, dal momento che per ipotesi $E(u|x_1,\dots,x_k)=0$, $Var(u|x)=E(u^2|x)$, e quindi l'ipotesi di varianza costante diventa $E(u^2|x_1,\dots,x_k)=E(u^2)=\sigma^2$. Il test di Breusch-Pagan si svolge come segue. Sotto le condizioni:

$$H_0 = la\ varianza \text{ è } costante$$

$$H_1 = H_0 \text{ è } falsa$$

Si ricavano gli errori di questa regressione (e_i) come stima di u_i e si elevano al quadrato, per stimare la varianza. A questo punto, si stima la seguente regressione:

$$e_i^2 = \beta_0 + \beta_1 \hat{y}_i$$

Dove $\hat{y}_i$, che rappresenta il valore di y_i stimato dalla regressione, viene usato per vedere se e^2 è correlato alle variabili dell'equazione iniziale. Si prende lo R^2 di questa regressione, $R^2_{BP.}$ Sotto le ipotesi sopra indicate, la variabile

$$\chi = n \cdot R^2_{BP}$$

segue asintoticamente la distribuzione χ^2_1. Se il valore *p* associato risulta essere inferiore al livello di significatività scelto, si può rifiutare l'ipotesi di omoschedasticità. (Wooldridge 2003)

Il test di White parte dalla stessa osservazione sull'uso di e^2 come stima della varianza della regressione, e prevede, sotto le stesse ipotesi:

$$H_0 = la\ varianza \text{ è } costante$$

$$H_1 = H_0 \text{ è } falsa$$

di regredire e^2 su tutte le variabili esplicative della regressione originaria, sui loro valori al quadrato, e su tutti i prodotti incrociati. Intuitivamente, l'ipotesi di varianza costante regge

solo se tutti i parametri di questa regressione sono uguali a zero, ovvero se R^2 di questa regressione (R^2_w) è uguale a 0. Per testarlo formalmente, sotto le ipotesi elencate la variabile

$$\chi = n \cdot R_W^2$$

segue asintoticamente la distribuzione χ_k^2, dove *k* rappresenta il numero di regressori dell'equazione. Se il valore *p* associato risulta essere inferiore al livello di significatività scelto, si può rifiutare l'ipotesi di omoschedasticità. (Gujarati & Porter 2010)

I risultati dei due test sono elencati nella tabella seguente, e suggeriscono entrambi di rifiutare l'ipotesi di omoschedasticità.

Breusch-Pagan	chi^2(1) = 190.03	Prob > chi^2 = 0.0000
White	chi^2(101) = 440.67	Prob > chi^2 = 0.0000

Risulta di conseguenza necessario ricorrere alla stima degli errori standard robusti, ovvero consistenti anche in presenza di eteroschedasticità, secondo il metodo suggerito da White (si veda Wooldridge 2003). In questo modo è infatti possibile calcolare le statistiche *t* e *F* in modo corretto.

LNHOURPAY	**Coef.**	**Robust Std. Err.**	**t**	**P>|t|**	**95% Conf. Interval**	
Constant	1,350416	.0433584	31.15	0.000	1,265424	1,435408
YEDUC	.0508385	.0027111	18.75	0.000	.0455242	.0561528
PTIME	-.2807831	.0135836	-20.67	0.000	-.3074097	-.2541564
PUBLIC	.1095704	.0118288	9.26	0.000	.0863834	.1327574
FEMALE	-.1310353	.0116975	-11.20	0.000	-.1539649	-.1081058
POTEXP	.0312425	.001516	20.61	0.000	.0282708	.0342142
POTEXP2	-.0004849	.0000295	-16.43	0.000	-.0005427	-.000427
TRAINING	.0710621	.011906	5.97	0.000	.0477238	.0944004
LONDON	.2588164	.0200689	12.90	0.000	.219477	.2981559
SCOTLAND	.0618323	.019665	3.14	0.002	.0232847	.1003799
SOUTH	.0955729	.0124024	7.71	0.000	.0712614	.1198843
COHABITANT	.1701433	.0115699	14.71	0.000	.1474639	.1928227
BLACK	-.2215648	.0350623	-6.32	0.000	-.2902944	-.1528353
ASIAN	-.239672	.0349135	-6.86	0.000	-.3081099	-.171234
INDIAN	-.0997315	.0346034	-2.88	0.004	-.1675616	-.0319014

R-squared	0.2442
F(14, 9503)	212.50
Prob > F	0.0000

Anche con gli standard errors robusti, tutti i coefficienti risultano significativi persino al livello dell'1%.

1.4 Errata specificazione del modello: il test di Ramsey - RESET

Dal momento che il modello economico dell'istruzione suggerisce che i risultati di questa regressione possano essere inficiati dall'omissione di una variabile rilevante, è necessario stabilire se questo sia il caso o no. J. Ramsey ha sviluppato un test generale sull'errata specificazione del modello (test di Ramsey – RESET) che parte dall'idea che l'errata specificazione di un modello, ad esempio a causa dell'omissione di una variabile, provoca un certo legame tra il termine di errore e la variabile dipendente stimata. Questo suggerisce che se inserendo $\hat{Y}$ in qualche forma tra le variabili dipendenti ($\hat{Y}^2, \hat{Y}^3, ...$) si ottiene un aumento di R^2, ciò può indicare un'errata specificazione del modello. Formalmente, sotto le ipotesi:

$$H_0: il\ modello \text{ è } correttamente\ specificato$$

$$H_1: H_0 \text{ è } falsa$$

si ottiene $\hat{Y}_i$, ovvero Y stimata dal modello scelto. Dopodiché, si stima nuovamente il modello aggiungendo potenze di $\hat{Y}_i$ tra le variabili esplicative. Chiamando R^2_{OLD} e R^2_{NEW} gli R^2 di questa regressione, e k il numero di parametri nel nuovo modello, la variabile

$$F = \frac{(R^2_{NEW} - R^2_{OLD})/n°\ nuovi\ regressori}{(1 - R^2_{NEW})/(n-k)}$$

segue la distribuzione F con (n° nuovi regressori, (n-k)) gradi di libertà. Se la variabile è statisticamente significativa al livello scelto, il test suggerisce la presenza di una variabile endogena. (Gujarati & Porter 2010)

Purtroppo, come mostra la tabella sottostante con i valori della statistica ricavata applicando il test al modello in esame, questo sembra essere proprio il caso.

F(3, 9500)	32.43
Prob > F	0.0000

Date le gravi conseguenze che l'omissione di una variabile rilevante e correlata con una variabile esplicativa comporta, è necessario tentare di stimare il rendimento dell'istruzione attraverso l'uso di variabili strumentali.

2 – Il metodo delle variabili strumentali

Oltre al metodo dei minimi quadrati a due stadi (2SLS), per stimare il modello è stato usato anche il metodo "Limited information maximum likelihood" (LIML), che qui viene presentato senza spiegazione: esso è infatti considerato più robusto alla non rilevanza degli strumenti, e permette di ottenere una seconda stima del rendimento dell'istruzione.

2.1 Minimi quadrati a due stadi

I risultati ottenuti sono indicati nelle tabelle seguenti. Si noti che il coefficiente di determinazione non è stato riportato, dal momento che in questo caso esso perde di significato. Infatti in caso di correlazione tra *x* e *u* non è possibile scomporre la varianza di *Y* in $Var(\hat{Y}) + Var(u)$, pertanto la formula standard per la derivazione di R^2 (*ESS/TSS*) non ha un'interpretazione naturale (Wooldridge 2003).

LNHOURPAY	Coef.	Std. Err.	z	P>|z|	95% Conf.Interval	
Constant	.6154482	.8401973	0.73	0.464	-1,031308	2,262205
YEDUC	.100489	.0567433	1.77	0.077	-.0107259	.2117039
PTIME	-.2897568	.0167624	-17.29	0.000	-.3226104	-.2569032
PUBLIC	.0716664	.0451031	1.59	0.112	-.016734	.1600669
FEMALE	-.1227868	.0152229	-8.07	0.000	-.1526231	-.0929505
POTEXP	.0345638	.0041369	8.36	0.000	.0264557	.0426719
POTEXP2	-.0004479	.0000529	-8.47	0.000	-.0005515	-.0003442
TRAINING	.0489417	.028146	1.74	0.082	-.0062235	.1041068
LONDON	.1973724	.0728911	2.71	0.007	.0545085	.3402363
SCOTLAND	.0346851	.0374959	0.93	0.355	-.0388056	.1081758
SOUTH	.0761559	.0254625	2.99	0.003	.0262503	.1260614
COHABITANT	.1388943	.0375846	3.70	0.000	.0652298	.2125587
BLACK	-.2618201	.0591359	-4.43	0.000	-.3777243	-.145916
ASIAN	-.2997211	.0766967	-3.91	0.000	-.4500439	-.1493982
INDIAN	-.1723292	.0902365	-1.91	0.056	-.3491894	.004531
F(14, 9503)	160.66					
Prob > F	0.0000					

Prima di analizzare il parametro di YEDUC, è bene notare che il segno degli altri parametri è lo stesso prodotto con il metodo OLS. Per quanto riguarda il rendimento dell'istruzione, esso sembra sottostimato dall'OLS: il metodo dei 2SLS infatti produce una stima pari al 10%. Questo risultato appare in contraddizione con il modello dell'istruzione esposto nel primo capitolo, e potrebbe suggerire che individui con minore abilità tendano a studiare di più per aumentare la loro produttività e il proprio reddito, altrimenti troppo basso. È importante però sottolineare che si tratta solo di una stima: è necessario perciò guardare alla significatività dello stimatore e all'intervallo di confidenza.

2.1.1 Test di significatività congiunta e individuale

Per quanto riguarda la significatività congiunta dei parametri, il valore di 160 della variabile F permette di rifiutare con relativa certezza l'ipotesi nulla di non significatività. Per quanto riguarda la significatività individuale invece la statistica di riferimento segue approssimatamente, sotto l'ipotesi nulla che il rispettivo parametro è uguale a 0, la

distribuzione normale standard (ovvero con media 0 e varianza 1) (Cappuccio & Orsi 1995). L'interpretazione del valore P e dell'intervallo di confidenza tuttavia rimane la medesima. Questa analisi rivela il fatto che il numero di parametri individualmente significativi si riduce: la stessa YEDUC presenta un valore P del 7,7%, rendendola significativa al livello del 90%, ma non del 95%. Ciò significa che l'intervallo di confidenza a questo livello di significatività comprende anche lo 0, e a maggior ragione la stima ottenuta con il metodo OLS: la conclusione affrettata che il rendimento dell'istruzione è superiore a quanto appare con una semplice regressione lineare non è giustificata.

Il fatto che il metodo delle variabili strumentali produca intervalli di confidenza più ampi non giunge inaspettato, poiché deriva direttamente dalla varianza degli stimatori, più ampia rispetto a quella del metodo OLS.

Per avere più chiarezza sulla bontà e convenienza dell'uso di variabili strumentali è necessario testare le ipotesi chiave del metodo delle variabili strumentali, ovvero l'incorrelazione tra gli strumenti e l'errore dell'equazione strutturale e la correlazione tra la variabile endogena e gli strumenti esclusi.

2.1.2 Il test di sovraidentificazione di Sargan

Questo test verifica se tutti gli strumenti, esclusi e inclusi, sono effettivamente incorrelati con il termine di errore dell'equazione strutturale. Il procedimento è come segue:

$$H_0: E(z_j u) = 0 \; per \; ogni \; j$$

$$H_1: H_0 \text{ è } falsa$$

Si stima l'equazione strutturale con il metodo dei 2SLS, e si ottengono gli errori e_i. A questo punto, si regredisce e_i su tutti gli strumenti tramite OLS, e si ottiene il coefficiente di determinazione R^2. Sotto l'ipotesi nulla, la statistica:

$$n \cdot R^2$$

segue la distribuzione $\chi^2(q)$, dove q=(*n° strumenti esclusi – n° variabili endogene).* Come si vede, il test si può applicare se e solo se il modello è sovraidentificato, ovvero se e solo se $q>0$. Se il valore P risultante è significativo al livello scelto di significatività, si deve rifiutare l'ipotesi di esogeneità degli strumenti. (Wooldridge 2003; Cappuccio & Orsi 1995)

I risultati del test applicato al modello sono riportati nella tabella sottostante. Sebbene il valore P non sia significativo al livello del 5%, il suo valore molto vicino alla soglia critica suggerisce che gli strumenti scelti potrebbero non essere buoni.

Chi-sq(3)	7.592
P-val	0.0552

2.1.3 Correlazione tra variabile endogena e strumenti esclusi

Per testare questa ipotesi, come già suggerito, basta verificare se almeno uno tra i coefficienti degli strumenti esclusi è individualmente significativo nella regressione di YEDUC su tutti gli strumenti. La tabella mostra i risultati di questa regressione: come si vede, solo uno strumento (HINDU), risulta significativo al 5% di confidenza, mentre MUSLIM e soprattutto BUDHIST presentano un valore P molto alto. L'interpretazione che scaturisce da questi risultati è che la bassa correlazione tra la variabile endogena YEDUC e gli strumenti esclusi potrebbe portare al problema degli strumenti deboli.

YEDUC	**Coef.**	**Std. Err.**	**z**	**P>|z|**	**95% Conf. Interval**	
Constant	1.480623	.1100433	134.55	0.000	1.459055	1.502191
MUSLIM	-.3105639	.4037528	-0.77	0.442	-1.101905	.480777
HINDU	.767087	.3714299	2.07	0.039	.0390978	1.495076
BUDHIST	-.1430124	.6745987	-0.21	0.832	-1.465202	1.179177
OTHER	.3630453	.2324676	1.56	0.118	-.0925828	.8186734
PTIME	.1837814	.0752601	2.44	0.015	.0362743	.3312886
PUBLIC	.7640052	.0670679	11.39	0.000	.6325546	.8954558
FEMALE	-.1701102	.0645163	-2.64	0.008	-.2965597	-.0436607
POTEXP	-.0668469	.0094171	-7.10	0.000	-.0853041	-.0483897
POTEXP2	-.000749	.0001803	-4.16	0.000	-.0011023	-.0003957
TRAINING	.4448255	.0675069	6.59	0.000	.3125143	.5771367
LONDON	1.219279	.1193657	10.21	0.000	.9853268	1.453232
SCOTLAND	.5395442	.1144668	4.71	0.000	.3151933	.763895
SOUTH	.3844201	.0639854	6.01	0.000	.2590111	.5098291
COHABITANT	.6277413	.062925	9.98	0.000	.5044106	.751072
BLACK	.8462741	.2958853	2.86	0.004	.2663496	1.426199
ASIAN	1.330029	.3540437	3.76	0.000	.6361163	2.023942
INDIAN	1.039507	.2834141	3.67	0.000	.4840259	1.594989

2.1.4 Errori standard robusti all'eteroschedasticità

Come per i minimi quadrati ordinari, anche per il metodo dei minimi quadrati a due stadi è possibile ottenere stime dell'errore standard dei coefficienti robuste all'eteroschedasticità. I risultati sono elencati nella tabella sottostante.

LNHOURPAY	**Coef.**	**Std. Err.**	**z**	**P>|z|**	**95% Conf. Interval**	
Constant	0,6154482	0,9080578	0,68	0,498	-1,1643120	2,3952090
YEDUC	0,1004890	0,0613305	1,64	0,101	-0,0197166	0,2206947
PTIME	-0,2897568	0,0185485	-15,62	0,000	-0,3261112	-0,2534024
PUBLIC	0,0716664	0,0482956	1,48	0,138	-0,0229911	0,1663240
FEMALE	-0,1227868	0,0157248	-7,81	0,000	-0,1536068	-0,0919668
POTEXP	0,0345638	0,0044088	7,84	0,000	0,0259228	0,0432048

POTEXP2	-0,0004479	0,0000550	-8,14	0,000	-0,0005557	-0,0003400
TRAINING	0,0489417	0,0295139	1,66	0,097	-0,0089045	0,1067878
LONDON	0,1973724	0,0779571	2,53	0,011	0,0445793	0,3501656
SCOTLAND	0,0346851	0,0396026	0,88	0,381	-0,0429345	0,1123047
SOUTH	0,0761559	0,0271352	2,81	0,005	0,0229718	0,1293399
COHABITANT	0,1388943	0,0402025	3,45	0,001	0,0600989	0,2176896
BLACK	-0,2618201	0,0675347	-3,88	0,000	-0,3941857	-0,1294546
ASIAN	-0,2997211	0,0808660	-3,71	0,000	-0,4582155	-0,1412266
INDIAN	-0,1723292	0,0951163	-1,81	0,070	-0,3587538	0,0140954

F(14, 9503)	160.66
Prob > F	0.0000

La maggior parte dei parametri vedono il proprio valore P aumentare leggermente. In particolare, il valore P di YEDUC aumenta al 10.1%, con un conseguente allargamento dell'intervallo di confidenza al 95%.

Il test di sovraidentificazione, che nel caso di presenza di eteroschedasticità prende il nome di test di Hansen, è però migliore: il valore P della statistica è infatti del 13.37%.

Chi-sq(3)	5.584
P-val	0.1337

2.1.5 Test di endogeneità di Wu-Hausman

Come ultimo test, si presenta un test per l'endogeneità della variabile YEDUC. Come infatti si è detto in precedenza, anche se i risultati ottenuti con il metodo dei 2SLS rimangono validi, in caso di esogeneità della variabile YEDUC è preferibile usare il metodo OLS che risulta più efficiente. Questo test, si svolge sotto l'ipotesi nulla che la variabile di interesse sia esogena, ovvero che $E(x_i \cdot u_i) = 0$. La statistica segue la distribuzione chi^2 con gradi di libertà pari al numero di variabili testate (Cappuccio & Orsi 1995). I risultati del test nel caso in esame sono esposti nella tabella. Dal momento che il test non è significativo al livello, ad esempio, del 5% o 10%, non si può rifiutare l'ipotesi nulla di considerare la variabile YEDUC come esogena. I risultati di questo test sembrano dunque suggerire l'utilizzo del metodo OLS per stimare il modello.

Chi-sq(1)	0.691
P-val	0.4058

2.2 Limited information maximum likelihood

Questo metodo di stima, proposto nel 1949 da Anderson e Rubin, è considerato più robusto alla non rilevanza degli strumenti. Per questo motivo, si è scelto di presentare i risultati ottenuti con questo metodo, senza tuttavia spiegarlo.

LNHOURPAY	Coef.	Std. Err.	z	P>|z|	95% Conf. Interval	
Constant	-0,54082	2,941202	-0,18	0,854	-6,30547	5,223833
YEDUC	0,1786	0,19868	0,90	0,369	-0,21081	0,568006
PTIME	-0,30387	0,040905	-7,43	0,000	-0,38405	-0,2237
PUBLIC	0,012035	0,152052	0,08	0,937	-0,28598	0,310052
FEMALE	-0,10981	0,035737	-3,07	0,002	-0,17985	-0,03977
POTEXP	0,039789	0,013451	2,96	0,003	0,013426	0,066152
POTEXP2	-0,00039	0,000152	-2,56	0,010	-0,00069	-9,2E-05
TRAINING	0,014141	0,088834	0,16	0,874	-0,15997	0,188254
LONDON	0,100708	0,247369	0,41	0,684	-0,38413	0,585541
SCOTLAND	-0,00802	0,111684	-0,07	0,943	-0,22692	0,210873
SOUTH	0,045609	0,07879	0,58	0,563	-0,10882	0,200035
COHABITANT	0,089733	0,125256	0,72	0,474	-0,15576	0,33523
BLACK	-0,32515	0,179238	-1,81	0,070	-0,67645	0,02615
ASIAN	-0,39419	0,242288	-1,63	0,104	-0,86907	0,080684
INDIAN	-0,28654	0,291696	-0,98	0,326	-0,85826	0,285173

F(14, 9503)	110.86
Prob > F	0.0000

I risultati ottenuti con questo metodo sono peggiori di quanto si ottiene con il metodo 2SLS: molti meno coefficienti sono significativi, e gli intervalli di confidenza sono decisamente più ampi. In particolare, il parametro di YEDUC ha un valore P del 37%, che non permette di rifiutare l'ipotesi nulla che esso sia uguale a 0, nonostante produca una stima del rendimento dell'istruzione elevatissima, pari al 17.86%.

Il test di Hansen per l'esogeneità degli strumenti, tuttavia, riporta un valore P molto elevato, che permette di rifiutare l'ipotesi nulla che gli strumenti esclusi siano correlati con l'errore strutturale.

Chi-sq(3)	3.458
P-val	0.3263

Il problema degli strumenti deboli

I risultati ottenuti col metodo delle variabili strumentali – sia attraverso il metodo 2SLS che il metodo LIML – perdono di significato se il modello soffre del cosiddetto problema degli strumenti deboli. Spiegato in modo informale, questo problema può sorgere nel momento in cui la correlazione tra variabile endogena e strumenti esclusi non è elevata e allo stesso tempo questi strumenti e il termine di errore dell'equazione strutturale presentano un grado, anche basso, di correlazione. Se questo accade, gli stimatori 2SLS (e LIML) possono risultare distorti persino più dei corrispettivi OLS. Dal momento che la correlazione tra strumenti e YEDUC nel modello non appare elevata, si ritiene necessario definire in modo più formale il problema e verificarne la sussistenza attraverso un test.

In precedenza è stato mostrato il valore dell'errore dello stimatore IV, che in presenza di correlazione tra lo strumento e l'errore strutturale risulta pari a (Wooldridge 2003):

$$plim\hat{\beta}_{1,IV} = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

Un altro modo di scrivere il secondo termine, ovvero la differenza tra $\hat{\beta}_{1,IV}$ e β_1, è il seguente (Cappuccio & Orsi 1995):

$$E(\hat{\beta}_{IV} - \beta) = \frac{\sigma_{u,v}}{\sigma_v^2}\left[\frac{1}{F+1}\right]$$

Dove *u* e *v* sono gli errori, rispettivamente, dell'equazione strutturale e di quella del primo stadio, mentre *F* è il valore della statistica del test F nell'equazione del primo stadio. Da questa espressione si ricavano le seguenti conclusioni (Cappuccio & Orsi 1995):

- Se F è grande, la distorsione tende a essere piccola, e dunque lo stimatore IV produce buoni risultati.
- Se F=0, si ha che la distorsione del metodo IV e OLS coincidono: infatti con questo valore di F, *x* = *v*, e l'espressione si può scrivere come $\frac{\sigma_{u,x}}{\sigma_x^2}$, ovvero esattamente la distorsione dello stimatore OLS.
- Se F è "piccolo", la distorsione tende a centrarsi intorno a $\frac{\sigma_{u,v}}{\sigma_v^2}$.

Anche se non è facile stabilire cosa si intenda per "piccolo", questo fatto causa numerosi problemi, ovvero (Cappuccio & Orsi 1995):

- Lo stimatore ha una distorsione elevata
- La sua distribuzione campionaria non approssima la normale, con la conseguenza che tutte le statistiche prodotte sulla base del suo valore perdono di significato, così come gli intervalli di confidenza.
- Il problema non riguarda solo campioni finiti o piccoli, ma è presente anche asintoticamente.

È bene notare che aggiungendo strumenti nell'equazione di primo stadio con una bassa correlazione con *x* non risolve il problema, anzi, dal momento che la statistica *F* dipende inversamente dal numero di variabili esplicative inserite (Cappuccio & Orsi 1995).

1 Il test per gli strumenti deboli di Stock e Yogo

Partendo dalle constatazioni precedenti sul ruolo della statistica *F* nel segnalare la debolezza o meno degli strumenti, Stock e Yogo propongono un test per determinarne la rilevanza diviso in due parti (per una formulazione più precisa e formale, si veda Cappuccio & Orsi 1995). La prima parte riguarda l'ipotesi nulla che la distorsione di $\hat{\beta}_{IV}$ sia maggiore, al 5% di significatività, di una certa percentuale rispetto alla distorsione di β_{OLS} (ad esempio il 10%). Questo sistema di ipotesi è chiamato di "distorsione relativa". La seconda parte del test invece parte dalla constatazione che, volendo testare la significatività del parametro della variabile endogena nell'equazione strutturale, se gli strumenti sono deboli la statistica $\frac{\hat{\beta}}{se_{\hat{\beta}}}$ ha una distribuzione che non approssima, neanche asintoticamente, la normale standard. Di conseguenza, se il test porta ad accettare un livello di significatività ad esempio del 5%, il livello di significatività reale potrebbe essere molto diverso. Le ipotesi da testare sono quindi le seguenti: scelto un valore di significatività nominale (es. 5%), la dimensione reale è maggiore del 10%, 15%, 20%, eccetera. Scelto il valore critico di riferimento, se la statistica è maggiore di quella critica, si può rifiutare l'ipotesi di strumenti deboli. Questo sistema di ipotesi è chiamato di "dimensione effettiva". Si noti che il valore della statistica per rifiutare l'ipotesi nulla della dimensione effettiva è maggiore rispetto a quello richiesto dal sistema di ipotesi della distorsione relativa. Nella tabella seguente sono elencati i valori critici del test per vari valori di distorsione relativa e dimensione effettiva nel caso di 4 strumenti esclusi.

5% maximal IV relative bias	16.85	**10% maximal IV size**	24.58	**10% max LIML size**	5.44
10% maximal IV relative bias	10.27	**15% maximal IV size**	13.96	**15% max LIML size**	3.87
20% maximal IV relative bias	6.71	**20% maximal IV size**	10.26	**20% max LIML size**	3.30
30% maximal IV relative bias	5.34	**25% maximal IV size**	8.31	**25% max LIML size**	2.98

La statistica F ottenuta è uguale a 2.801: dal momento che è inferiore a qualsiasi valore critico, le ipotesi nulle di presenza di strumenti deboli non possono essere rifiutate. Questo risultato è molto pesante, perché indica l'inconsistenza della stima del rendimento dell'istruzione ottenuta tramite i metodi delle variabili strumentali, oltre all'inaffidabilità dei test prodotti sulla base di questo parametro. Ciò significa che oltre ai test di significatività, non sono affidabili neppure il test di Sargan sulla validità degli strumenti e il test di endogeneità di Hausman, che aveva suggerito, in contrasto col modello economico, l'esogeneità della variabile YEDUC. La conclusione che se ne trae è che gli strumenti scelti non permettono una stima ragionevole del rendimento dell'istruzione, e non possono quindi essere usati per tale scopo.

Conclusione

Dopo aver esposto la teoria economica sottostante al modello, vale a dire il modello dell'istruzione, e le basi econometriche sulle quali è stata costruita l'analisi, ovvero i metodi di regressione dei minimi quadrati ordinari e delle variabili strumentali, si è passati all'illustrazione delle variabili scelte e della forma funzionale adottata. La parte più importante è però relativa ai risultati ottenuti tramite i vari metodi. Nella tabella sottostante sono indicati i rendimenti marginali dell'istruzione stimati attraverso i vari metodi. Questi rendimenti sono calcolati direttamente dal valore del parametro assegnato a YEDUC, riportato insieme al valore P del relativo test di significatività nelle ultime due colonne.

	R. Marg. Istruzione	β	Valore P
OLS	5.08%	0.0508385	0.000
OLS robusti	5.08%	0.0508385	0.000
2SLS	10.05%	0.1004890	0.077
2SLS robusti	10.05%	0.1004890	0.101
LIML (robusti?)	17.86%	0.1786000	0.369

Mentre il metodo dei minimi quadrati ordinari suggerisce un rendimento dell'istruzione attorno al 5%, i due metodi delle variabili strumentali indicano un rendimento marginale compreso che varia tra il 10% e il 18%. Tuttavia, mentre la stima OLS è significativamente diversa da 0, anche correggendo per l'eteroschedasticità, lo stesso non si può dire per le stime prodotte col metodo 2SLS e, soprattutto, LIML.

Nonostante le considerazioni fatte in merito alla variabile esclusa "abilità" suggeriscano di utilizzare delle variabili strumentali per produrre una stima corretta del rendimento, il test sugli strumenti deboli indica come quelli scelti, ovvero i credi religiosi, non siano adatti allo scopo e producano risultati fortemente distorti. Per poter ottenere risultati validi, è necessario quindi scegliere nuovi strumenti, o in alternativa trovare variabili correlate con l'abilità da usare come proxy per la stessa.

Bibliografia

Ashenfelter, O.C. & Krueger, A.B., 1994. Estimates of the Economic Return to Schooling from a New Sample of Twins. *American Economic Review*, (84), pagg.1157–1173.

Borjas, G.J., 2013. *Labor Economics* 6th ed. McGraw-Hill, a c. di, New York: McGraw-Hill.

Cappuccio, N. & Orsi, R., 1995. *Econometria*, Il Mulino.

Card, D., 2001. Estimating the return to schooling: progress on some persistent econometric problems. *Econometrica*, 69(5).

Gujarati, D.N. & Porter, D.C., 2010. *Essentials of Econometrics* 4th ed., McGraw-Hill.

Leigh, A. & Ryan, C., 2008. Estimating returns to education using different natural experiment techniques. *Economics of Education Review*, (27), pagg.149–160.

Office for national statistics, 2016. *Labour force survey*, Available at: https://discover.ukdataservice.ac.uk/series/?sn=2000026.

Psacharapoulos, G., 1985. Returns to Education: A Further International Update and Implications. *Journal of Human Resources*, (20), pagg.583 – 604.

Taubman, P., 1976. Earnings, Education, Genetics, and Environment. *Journal of Human Resources*, (11), pagg.447–461.

Treccani, 2012. Economia politica. *Dizionario di economia e finanza*.

Trostel, P., Walker, I. & Woolley, P., 2002. Estimates of the economic return to schooling for 28 countries. *Labour Economics*, 9(1), pagg.1–16.

Wooldridge, J.M., 2003. Introductory Econometrics: A Modern Approach. *Economic Analysis*, 2nd.